
Guided Policy Search with Delayed Sensor Measurements

Connor Schenck and Dieter Fox

Department of Computer Science & Engineering, University of Washington

Abstract

Guided policy search [1] is a method for reinforcement learning that trains a general policy for accomplishing a given task by guiding the learning of the policy with multiple guiding distributions. Guided policy search relies on learning an underlying dynamical model of the environment and then, at each iteration of the algorithm, using that model to gradually improve the policy. This model, though, often makes the assumption that the environment dynamics are markovian, e.g., depend only on the current state and control signal. In this paper we apply guided policy search to a problem with non-markovian dynamics. Specifically, we apply it to the problem of pouring a precise amount of liquid from a cup into a bowl, where many of the sensor measurements experience non-trivial amounts of delay. We show that, with relatively simple state augmentation, guided policy search can be extended to non-markovian dynamical systems, where the non-markovianity is caused by delayed sensor readings.

1 Introduction

Reinforcement learning is a well studied problem in machine learning [2]. Many researchers have successfully applied various reinforcement learning methods to different virtual domains [3, 4, 5, 6]. While there have been applications of reinforcement learning to the physical domain on real robots, they are usually very limited in scope [7, 8]. More recently, work by Levine *et al.* [1, 9] has shown how reinforcement learning can be applied in more unstructured robotic applications. Still, the task of training policies in real-world domains on physical robots remains a challenging one.

In this paper, we apply reinforcement learning to a real-world task with non-trivial sensor delays. Specifically, we look at using guided policy search to learn a policy for pouring water. The goal of the robot is to pour a precise amount of liquid from a cup into a bowl. The delay is introduced by both the time the water takes to reach the bowl and the scale sensing the weight change of the bowl. The precise nature of the pouring task necessitates that the robot have at least a cursory understanding of the fluid dynamics involved, a highly non-trivial problem.

To solve this task, the robot first trains a dynamics model from a set of example trajectories. Then, given this dynamics model, it alternates between a single step of trajectory optimization and updating the weights of a neural network to match the optimized trajectories. Next, the robot rolls out the trained neural network on the real system, using the resulting sensor data to retrain the dynamics model. This process repeats until the robot converges. In order for the robot to reason with delayed sensor measurements, it uses the previous n states ($n > 1$), rather than just the previous state, to predict both the dynamics and policy controls.

We show that a robot can successfully utilize this methodology to learn a policy enabling it to pour a precise amount of liquid. We varied the initial amount of water in the container, and for each initial amount, the robot attempted to pour a precise amount into the bowl. The robot converged after approximately 25 iterations. It was able to pour within 10g of the target for all initial conditions.

The rest of this paper is organized as follows. The next section details relevant prior work to this paper. Section 3 describes our experimental setup. Section 4 lays out the methodology we employed to solve this task in detail. Section 5 describes how we evaluated the robot on the pouring task. Section 6 details the results the robot was able to achieve. And finally section 7 concludes the paper and describes some potential avenues for future work.

2 Related Work

Many various aspects of the task of robotic pouring have been investigated by prior work. Some studies have focused on utilizing specialized hardware and algorithms to achieve very precise pouring results [10], while others have focused on learning the broad motions of pouring through human demonstrations [11, 12]. Okada *et al.* [13] used a motion planner to manipulate a pouring object, Yamaguchi and Atkeson [14] used differential dynamic programming to pour in a simulator, and Tamosiunaite *et al.* used dynamic movement primitives to learn the goal and shape of a pouring trajectory, but all of these were designed to pour the entire contents of the container out, rather than a precise amount. To the authors’ knowledge, the only study to attempt to combine learning and precise pouring was done by Rozo *et al.* [15], who used human demonstrations to learn to pour 100 ml from a bottle into a cup.

However, there have been multiple studies applying reinforcement learning to other tasks on real robotic systems. Work by Konidaris *et al.* [16, 17, 18] has attempted to show how a robot can learn complex tasks by learning simpler skills and chaining them together. Further work by Niekum [19, 20] showed how a robot can learn complex multi-step tasks from unstructured demonstrations. But these works represented learning of rather imprecise, though complex, goals (e.g., pressing a large button to open a door). Work by Deisenroth *et al.* [21, 22] on the PILCO model-based reinforcement learning framework, though, has focused on learning more precise tasks such as the cart-pole task and the block stacking task. Indeed, the methodology used in this paper is similar to that in the PILCO framework, with the robot alternatively rolling out on the real system, training a dynamics model, and then fitting the policy parameters. A major difference between PILCO and our methods is that we empirically determined a time-based locally-linear dynamics model performs better for the pouring task than the gaussian process dynamics model used by PILCO.

Another major difference between PICLO and our methodology is that we use trajectory optimization to facilitate learning the policy parameters, rather than trying to fit them directly. We use iterative linear quadratic gaussians [23] in this paper, which is a trajectory optimization method based off of the linear quadratic regulator trajectory optimization method [24], both of which are types of differential dynamic programming [25]. Recently, work by Mordatch and Todorov [26] has shown how these methods can be applied to facilitate policy learning in a simulated environment, although this method is difficult to adapt to real environments where the dynamics are unknown.

Levine *et al.* [1, 9] have developed a methodology similar to [26] called guided policy search (GPS) that works on real robots in physical environments. Initially, they applied GPS only to simulated problems [1], but in follow-up work [9] they showed how it can be used to solve tasks on a real robot such as putting a cap on a bottle, inserting a brick into a block, and hanging a hanger on a rack. Our work in this paper is heavily inspired by the work of Levine *et al.*. Here, we apply GPS on a real robot in an environment with non-trivial sensor delay, specifically, to the task of pouring a precise amount of liquid into a bowl.

3 Experimental Setup

3.1 Robotic Platform

The robot used in this paper was the Rethink Robotics Baxter Research Robot, pictured in figure 1. It is an upper-torso humanoid robot with two 7-degree-of-freedom arms. Each arm is equipped with an electric parallel gripper. The motors on each joint can be controlled using position controls (via a built-in PID controller), velocity controls, or torque controls. In this paper, we controlled the

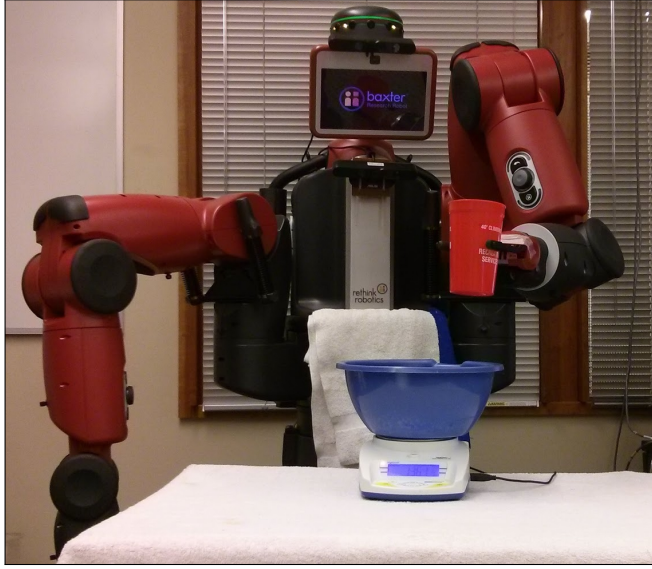


Figure 1: The robot used in these experiments. It is a Rethink Robotics Baxter Research Robot, equipped with two 7-DOF arms. Also shown is the experimental setup with the cup in the robot’s gripper positioned above the table, with the bowl resting on top of the scale.

robot’s arm using the velocity control mode. The joints are equipped with joint encoders and torque sensors¹. In the experiments in this paper, we used only the robot’s left arm.

3.2 Experimental Environment

The robot was placed in front of a small table. On the table was an Adam Equipment HCB 3000 Highland Portable Precision Balance scale. The scale had a maximum capacity of three kilograms and a resolution of one tenth of one gram. This scale was selected for its ability to provide real-time readings via USB cable to an attached computer. The scale has a built-in filter that introduces an approximately 0.5 to 1.0 second sensor delay in the measurements taken from the scale. On top of the scale was placed a medium sized plastic mixing bowl. A plastic cup was placed in the robot’s left gripper. This configuration is shown in figure 1. The cup was pre-filled by the experimenters with a precise amount of water between two-hundred and four-hundred grams.

3.3 Data Collection

We ran 500 pouring trials. Each trial began with the cup already in the robot’s gripper and pre-filled by the experimenter. The trial ended after exactly twenty-five seconds had passed, regardless of whether or not the robot had poured any liquid. During each trial, the robot’s joint angles, joint velocities, and the scale value were recorded at a rate of two hertz. While the robot is equipped with many other sensors, the robot only used its own joint angles, velocities, and the scale value to learn the pouring task.

3.4 State Space

In order to isolate the precise pouring task, in this paper, we fixed the robot’s wrist over the bowl on the table and gave it control only over its last joint (the wrist joint), effectively letting it control only the angle of the cup. The robot used a four-dimensional state space. The first dimension was the angle of the robot’s wrist in radians, shifted so that zero is when the cup is upright and π when the cup is inverted. The second dimension is the amount, in grams, remaining to pour until the robot reaches the pouring target. In this way, the target pouring amount is implicitly built into the

¹We empirically determined that the torque sensors built into the robot’s joints are not reliable and so did not use them in this paper.

state space. The third dimension is the change in the second dimension from the last timestep to the current one. And finally, the fourth dimension is the amount of water in the cup. This amount is initialized to a specific, known amount, and then updated throughout each trial by subtracting the change in the scale value.

4 Methodology

4.1 Problem Definition

Let $\mathcal{X} \in \mathbb{R}^d$ be the d dimensional state space the robot operates in, and let $\mathcal{X}_{init} \subseteq \mathcal{X}$ be the set of valid starting states. The goal of the robot is to learn a policy $\pi(\mathbf{x}_t; \theta) \rightarrow \mathbf{u}_t$ that minimizes a cost function l , where \mathbf{x}_t is the state at time t , \mathbf{u}_t is the robot's control signal at time t , and θ is the learned policy parameters. The robot must learn a policy that, from any initial state $x_1 \in \mathcal{X}_{init}$, generates a trajectory $(\mathbf{x}_1, \mathbf{u}_1), \dots, (\mathbf{x}_T, \mathbf{u}_T), (\mathbf{x}_{T+1})$ that minimizes $\sum_{i=1}^T [l(\mathbf{x}_i, \mathbf{u}_i)] + l(\mathbf{x}_{T+1})$ over a fixed horizon T .

4.2 Algorithm Overview

In this paper, we use a modified version of guided policy search [1] to train policy parameters θ . The robot is given N initial example trajectories $\tau = \{(\mathbf{x}_1, \mathbf{u}_1), \dots, (\mathbf{x}_T, \mathbf{u}_T), (\mathbf{x}_{T+1})\}^N$, where \mathbf{x}_t is the state of the robot at time t and \mathbf{u}_t is the control applied at time t . Next the robot trains a time-based dynamics model $\hat{f}_t(\mathbf{x}_t, \mathbf{u}_t) \rightarrow \mathbf{x}_{t+1}$. Using this model, the robot alternates between optimizing the trajectories τ using an iLQG backward pass and optimizing the policy parameters θ to match the updated trajectories. Next the robot uses the policy $\pi(\mathbf{x}_t; \theta) \rightarrow \mathbf{u}_t$ to rollout from each starting state \mathbf{x}_1 in each trajectory τ_i . Finally, the robot retrain the dynamics model \hat{f}_t and repeats this process until convergence.

4.3 Learning the Dynamics Model

We use a very similar learned dynamics model to [9]. The robot must learn the function $\hat{f}(\mathbf{x}_t, \mathbf{u}_t) \rightarrow \mathbf{x}_{t+1}$ mapping the state \mathbf{x}_t and control \mathbf{u}_t at the current timestep to the next state \mathbf{x}_{t+1} . The dynamics model is a time-based, locally linear model with a gaussian mixture model over the prior. Since the model is local, the robot learns a separate model for each each of the N example trajectories. The rest of this section will describe the training process for one model.

Given a training set $\tilde{\tau} = \{(\mathbf{x}_1, \mathbf{u}_1), \dots, (\mathbf{x}_T, \mathbf{u}_T), (\mathbf{x}_{T+1})\}^M$, where M is the number of iterations so far, the robot learns a separate linear function $\hat{f}_t(\mathbf{x}_t, \mathbf{u}_t) \rightarrow \mathbf{x}_{t+1}$ for all timesteps t . For applications on real robots, though, the number of training trajectories M can often be too small compared to the dimensionality of the state space to fit a linear model at each timestep, so instead the robot learns an equivalent model and uses shared dynamics between timesteps to compensate for the small amount of training data at each timestep.

At each timestep, the robot fits a multivariate gaussian $\mathcal{N}(\mu_t, \Sigma_t)$ over the concatenated vectors $\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle$ which we write as $\langle \mathbf{x}, \mathbf{u}, \mathbf{x}' \rangle$ for simplicity. To predict the next state after timestep t given \mathbf{x} and \mathbf{u} , the robot can simply condition the normal distribution on \mathbf{x} and \mathbf{u} and solve for \mathbf{x}' as follows:

$$\mathbf{x}' = \mu_{\mathbf{x}'} + \Sigma_{\mathbf{x}'(\mathbf{x}, \mathbf{u})} \Sigma_{\langle \mathbf{x}, \mathbf{u} \rangle \langle \mathbf{x}, \mathbf{u} \rangle}^{-1} (\langle \mathbf{x}, \mathbf{u} \rangle - \mu_{\langle \mathbf{x}, \mathbf{u} \rangle})$$

where μ_a is the components of μ who's elements pertain to a , and Σ_{ab} is the covariance between a and b extracted from Σ . Note that we only use the mean of the conditional distribution over \mathbf{x}' in this paper.

4.3.1 Estimating the Model Parameters

Let $\bar{\mu}$ and $\bar{\Sigma}$ be the empirical mean and covariance respectively for all $\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle$ for all t . The robot estimates an inverse-Wishart prior over the parameters of the T gaussian distributions

$\mathcal{N}(\mu_t, \Sigma_t)$. The prior has parameters Φ , μ_0 , m , and n . The parameters μ_t and Σ_t for the distribution at time t are estimated as follows:

$$\mu_t = \frac{m\mu_0 + M\hat{\mu}_t}{m + M} \quad \Sigma_t = \frac{\Phi + M\hat{\Sigma}_t + \frac{Mm}{n+m}(\hat{\mu}_t - \mu_0)(\hat{\mu}_t - \mu_0)^T}{M + n_0}$$

where $\hat{\mu}_t$ and $\hat{\Sigma}_t$ are the empirical mean and covariance respectively of the set $\{\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle\}^M$.

The prior parameters Φ , μ_0 , m , and n are estimated as

$$\Phi = n_0 \bar{\Sigma} \quad \mu_0 = \bar{\mu} \quad m = n_0 = 1.$$

4.3.2 Approximating Non-Linearity With a Gaussian Mixture Model

While the above method for estimating the gaussian parameters μ_t and Σ_t effectively reduces the number of training data points required at each time point t , the inverse-Wishart prior enforces a global linearity assumption on the dynamics model, as opposed to a local linearity assumption, which can make it difficult for the robot to operate in highly non-linear environments.

To handle this, the robot fits a gaussian mixture model [27] over the set $\{\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle\}_{t=1, \dots, T}^M$. Then, for each $t \in [1, T]$, the robot estimates the prior parameters Φ and μ_0 (n_0 and m remain constant at 1) as

$$\Phi = \frac{\sum_i p(\{\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle\}^M | \bar{\mu}_i, \bar{\Sigma}_i) \bar{\Sigma}_i}{\sum_i p(\{\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle\}^M | \bar{\mu}_i, \bar{\Sigma}_i)}$$

$$\mu_0 = \frac{\sum_i p(\{\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle\}^M | \bar{\mu}_i, \bar{\Sigma}_i) \bar{\mu}_i}{\sum_i p(\{\langle \mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1} \rangle\}^M | \bar{\mu}_i, \bar{\Sigma}_i)}$$

where $\bar{\mu}_i$ and $\bar{\Sigma}_i$ are the mean and covariance of the i th mixing element. Essentially, Φ and μ_0 are the weighted average of each mixing element. In this paper, we used the standard implementation of gaussian mixture models built into the Matlab Statistics and Machine Learning Toolbox [28].

4.4 Trajectory Optimization

Given a trained dynamics model \hat{f} and set of example trajectories τ , the robot must optimize those trajectories with respect to a given cost function l . However, if the robot uses l directly during trajectory optimization, the policy π may be unable to approximate the example trajectories. Instead the robot uses the modified cost function

$$l^*(\mathbf{x}, \mathbf{u}, \pi) = l(\mathbf{x}, \mathbf{u}) + \lambda \|\mathbf{u} - \pi(\mathbf{x})\|^2$$

where $\|x\|$ is the L2-norm and λ is the weight given to the second term of the equation. Informally, the second term of l^* enforces that whatever trajectory the optimizer finds, it should stay close to the policy.

4.4.1 iLQG Backward Pass

For each trajectory $\tau_i \in \tau$, the robot performs an iLQG backward pass to optimize the trajectory. It splits the combined optimization problem over $\mathbf{u}_1, \dots, \mathbf{u}_T$ into individual optimizations for each \mathbf{u}_t by optimizing the Q -function backwards in time, starting at time T . The Q -function is given by

$$Q(\delta \mathbf{x}_t, \delta \mathbf{u}_t) = l^*(\mathbf{x}_t + \delta \mathbf{x}_t, \mathbf{u}_t + \delta \mathbf{u}_t) + V_{t+1}(\hat{f}(\mathbf{x}_t + \delta \mathbf{x}_t, \mathbf{u}_t + \delta \mathbf{u}_t))$$

where $\delta \mathbf{x}_t$ and $\delta \mathbf{u}_t$ are the updates to apply to \mathbf{x}_t and \mathbf{u}_t respectively, and V_{t+1} is given by

$$V_{t+1}(\mathbf{x}) = \min_{\mathbf{u}} \left[l^*(\mathbf{x}, \mathbf{u}) + V_{t+2}(\hat{f}(\mathbf{x}, \mathbf{u})) \right].$$

The robot minimizes Q with respect to $\delta \mathbf{u}_t$ by taking the first and second derivatives of Q . V is intractable to differentiate directly, so the robot approximates V_{t+1} by substituting the next timestep's result, $Q(\delta \mathbf{x}_{t+1}, \delta \mathbf{u}_{t+1})$, in place of V_{t+1} . Thus it is necessary to work backwards through the trajectory so that $Q(\delta \mathbf{x}_{t+1}, \delta \mathbf{u}_{t+1})$ is already computed when computing $Q(\delta \mathbf{x}_t, \delta \mathbf{u}_t)$.

For further details of the iLQG algorithm, please refer to [23].

4.5 Policy Learning

Given the set of example trajectories $\tau = \{(\mathbf{x}_1, \mathbf{u}_1), \dots, (\mathbf{x}_T, \mathbf{u}_T), (\mathbf{x}_{T+1})\}^N$, fitting the policy parameters θ can be framed as a simple regression problem, i.e., train the parameters to map every $\mathbf{x}_t \rightarrow \mathbf{u}_t$. However, if N and T are small (e.g., 10 and 50, respectively), then that can leave relatively few training points for high-dimensional state spaces, which can make it difficult for the policy to smoothly fit a function over the uncovered areas of the space.

To solve this, the robot utilizes the gains matrices L_t generated during the iLQG backward pass to generate more training data points. The matrix L_t is used as follows

$$\mathbf{u}_t = \hat{\mathbf{u}}_t + L_t (\mathbf{x}_t - \hat{\mathbf{x}}_t)$$

where $\hat{\mathbf{x}}_t$ and $\hat{\mathbf{u}}_t$ are the open-loop trajectory found by iLQG at time t , and \mathbf{x}_t and \mathbf{u}_t are the actual state and controls when rolling out the trajectory on the real system. Intuitively, L_t computes how much to alter the open-loop control $\hat{\mathbf{u}}_t$ based on the difference between the open-loop state $\hat{\mathbf{x}}_t$ and the actual state \mathbf{x}_t . Thus, L_t makes the open-loop trajectory into a local policy around that trajectory.

To generate more data points, the robot draws many samples from a small gaussian around each $\mathbf{x}_t \in \tau$, and then used the corresponding gains matrix L_t to generate the controls for each sampled point. Adding these sampled data points to the original training data points, the robot can then formulate learning the policy parameters θ as a standard regression problem. For the experiments in this paper, the robot used a neural network to learn the policy.

Note that it is critical to the success of the robot to tightly interleave the trajectory optimization and policy learning iterations. That is, the robot does exactly one iLQG backward pass on each example trajectory, followed by updating the policy parameters θ to better fit the example trajectories. The robot repeats this inner loop multiple times before rolling out the policy in the real environment, and then returning to the inner loop. If the trajectory optimization and policy learning were not tightly interleaved, then the trajectory optimization could easily optimize the example trajectories in a way that would make it very difficult or impossible for the policy to learn. Instead, interleaving them in this manner keeps the policy “close” to the trajectory optimizer, and the policy deviation term in the cost function prevents the trajectory optimizer from moving the trajectories too far from what the policy can learn.

4.5.1 Training the Neural Network

The neural network layout used by the robot is shown in figure 2. The input \mathbf{x}_t is fed into a hidden layer with $|\mathbf{x}_t|$ hidden units. The output of each is the linear combination of its input fed through a rectified linear function, i.e.,

$$h_i(\mathbf{x}_t) = \max(\mathbf{x}_t \bullet \mathbf{w}_i, 0)$$

where \mathbf{w}_i is the set of weights for node i . Next, the output of the hidden layer is multiplied element-wise with the input. Finally, the result of the element-wise product is combined linearly into the output of the network.

The neural network used in this paper was implemented using the Caffe deep learning framework [29]. We used the built-in backpropagation to fit the weights of the network to the training data.

4.6 Handling Delay

Up to this point, we’ve described how the robot learns a policy as if there was no sensor delay. In order to learn in an environment with sensor delay, the robot augments its reasoning about states to include the previous n states. Thus, when learning the dynamics model, the function maps the previous n states and controls to the next state, i.e., $\hat{f}(\mathbf{x}_{t-n}, \mathbf{u}_{t-n}, \dots, \mathbf{x}_t, \mathbf{u}_t) \rightarrow \mathbf{x}_{t+1}$. Additionally, when learning the policy, it takes into account the previous n states as well, i.e., $\pi(\mathbf{x}_{t-n}, \dots, \mathbf{x}_t; \theta) \rightarrow \mathbf{u}_t$. So long as the previous n states cover the length of the delay, reasoning about them allows the robot to combine both delayed and non-delayed sensor readings to predict the next state or control.

5 Evaluation

We evaluated the robot on the pouring task. The robot’s arm was fixed over the bowl, and a cup was placed in its gripper. It was given control over its wrist joint so that it could control the angle of the

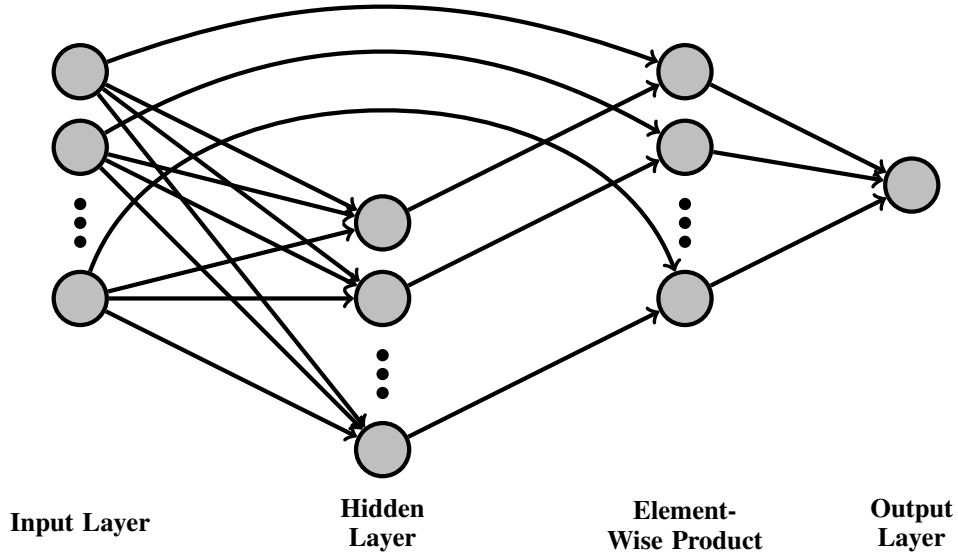


Figure 2: The neural network used to learn the policy. The input layer is fed into a fully-connected hidden layer, and then the output of the hidden layer is element-wise multiplied by the input layer, which is finally fed into the output layer.

cup. The robot controlled the angle by setting the angular velocity of its wrist joint. The cup was initialized upright before each trial and prefilled by the experimenter. The goal of the robot was to pour a specific amount of water into the bowl, and to be as accurate as possible.

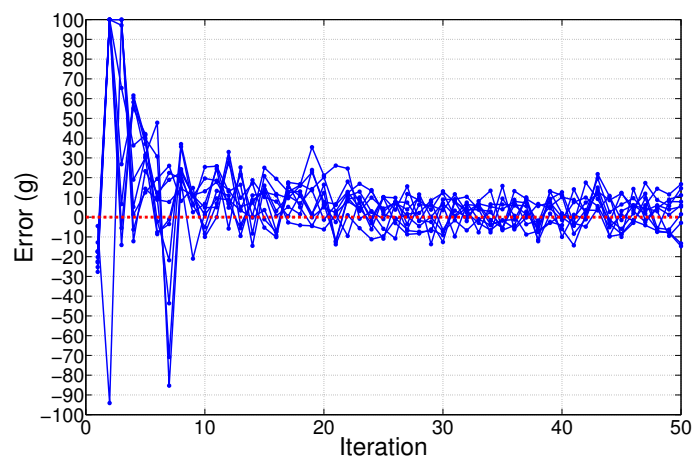
We set the number of example trajectories N to 10, and varied the initial amount of water in the cup for each trajectory uniformly between 200 and 400 grams, and we fixed the pouring target at 100 grams. The trajectories were initialized using a standard PID controller. We fixed the number of previous states to include n at 4 (i.e., the current state plus the 3 previous). The horizon for every trajectory was fixed to 50 timesteps, which, given a 2 hertz sampling rate, meant each trajectory lasted exactly 25 seconds. We chose to sample at 2 hertz to balance between having n high enough to cover the entire length of the sensor delay yet low enough to not cause the augmented state space to be too high-dimensional.

At the start of each iteration, the robot trained the dynamics model. Next, the robot alternated between one step of trajectory optimization and fitting the policy 10 times each. After finishing the inner loop of the algorithm, the robot rolled out each of the 10 trajectories from their starting states on the real system using the learned policy. Finally, the robot updated each example trajectory with the results of the real rollout, and then began the next iteration.

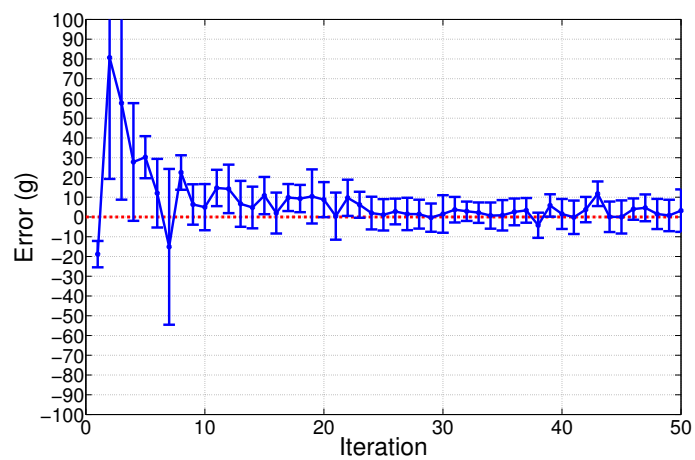
6 Results

The results are shown in figure 3. The error is reported in the number of grams of deviation the robot was from the target, that is, after a rollout, the difference between the number of grams of water in the bowl and the desired number, with values closer to 0 meaning better performance. Figure 3a shows the error for each example trajectory after each iteration. Figure 3b shows the mean and standard deviation of the example trajectories shown in figure 3a.

From the graphs, it is clear that the robot converged after approximately 25 iterations. Looking closely at figure 3a, it is apparent that iteration 33 was the first iteration where the robot was able to pour within 10 grams of the target for all trajectories. Furthermore, figure 3b shows that after iteration 23, the standard deviation of the robot’s error falls below 10 grams. In our experience, an accuracy of ± 10 grams is approximately what can be expected of a human performing the same task. Thus, the graphs in figure 3 show that the robot, after approximately 25 iterations, was able to converge to human-level performance.



(a) Error for each of the example trajectories



(b) Mean and standard deviation of the error

Figure 3: The error in grams after each iteration. The error is how far the robot was from the pouring target at the end of the rollout.

7 Conclusion and Future Work

In this paper, we used guided policy search to train a policy on a real robot to solve a task with non-trivial sensor delay. Specifically, the robot learned a policy for the pouring task. The goal of the robot was to pour a precise amount of water. It did this by iteratively pouring with its current policy, training a dynamics model, and then updating the policy using trajectory optimization. The robot was able to pour within 10 grams of the target (100 grams) after 33 iterations.

We showed that the robot was able to reach human-levels of performance on the pouring task. While this may not be high enough for tasks such as high precision manufacturing, it is sufficient for many household tasks such as cooking. Furthermore, we showed that, using a generic robotic platform, a robot can successfully learn to manipulate fluids. This is significant because, rather than relying on specialized hardware and algorithms, we showed that a generic learning platform can successfully be used to achieve human-level performance on a common household task.

This conclusion lends itself nicely to ideas for future work. Now that we know a robot can learn to pour as well as a human, in future work we could build on this by using relatively simple tasks like pouring to scaffold learning of more complex tasks such as multi-step cooking processes. Additionally, in future work can utilize the vast array of sensors available to modern robots to improve learning and foster a better understanding of the work environment.

References

- [1] Levine, S., Koltun, V.: Guided policy search. In: Proceedings of The 30th International Conference on Machine Learning. (2013) 1–9
- [2] Alpaydin, E.: Reinforcement learning. In: Introduction to machine learning. MIT Press (2014) 517–290
- [3] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. In: Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS) Workshop on Deep Learning. (2013)
- [4] Bowling, M., Veloso, M.: Multiagent learning using a variable learning rate. *Artificial Intelligence* **136**(2) (2002) 215 – 250
- [5] Riedmiller, M.: Neural fitted Q iteration—First experiences with a data efficient neural reinforcement learning method. In: Proceedings of the 16th European Conference on Machine Learning (ECML), Springer (2005) 317–328
- [6] McGovern, A., Barto, A.G.: Automatic discovery of subgoals in reinforcement learning using diverse density. In: Proceedings of the 18th International Conference on Machine Learning (ICML). (2001)
- [7] Smart, W.D., Kaelbling, L.P.: Effective reinforcement learning for mobile robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). (2002) 3404–3410
- [8] Kohl, N., Stone, P.: Policy gradient reinforcement learning for fast quadrupedal locomotion. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). (2004) 2619–2624
- [9] Levine, S., Finn, C., Darrell, T., Abbeel, P.: End-to-end training of deep visuomotor policies. In: Proceedings of the 32nd International Conference on Machine Learning (ICML) Workshop on Deep Learning. (2015)
- [10] Yano, K., Toda, T., Terashima, K.: Sloshing suppression control of automatic pouring robot by hybrid shape approach. In: Proceedings of the 40th IEEE Conference on Decision and Control (CDC). (2001) 1328–1333
- [11] Langsfeld, J., Kaipa, K., Gentili, R., Reggia, J., Gupta, S.: Incorporating failure-to-success transitions in imitation learning for a dynamic pouring task. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop on Compliant Manipulation: Challenges and Control, Chicago, Illinois, USA. (2014)

- [12] Cakmak, M., Thomaz, A.L.: Designing robot learners that ask good questions. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, ACM (2012) 17–24
- [13] Okada, K., Kojima, M., Sagawa, Y., Ichino, T., Sato, K., Inaba, M.: Vision based behavior verification system of humanoid robot for daily environment tasks. In: Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids). (2006) 7–12
- [14] Yamaguchi, A., Atkeson, C.G.: Differential dynamic programming with temporally decomposed dynamics. In: Proceedings of the IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). (2015) 696–703
- [15] Rozo, L., Jimenez, P., Torras, C.: Force-based robot learning of pouring skills using parametric hidden markov models. In: Proceedings of the IEEE 9th Workshop on Robot Motion and Control (RoMoCo). (2013) 227–232
- [16] Konidaris, G.: Autonomous Robot Skill Acquisition. PhD thesis, University of Massachusetts, Amherst (2011)
- [17] Konidaris, G., Kuindersma, S., Grupen, R., Barreto, A.S.: Constructing skill trees for reinforcement learning agents from demonstration trajectories. In: Advances in Neural Information Processing Systems 23. (2010) 1162–1170
- [18] Konidaris, G., Barreto, A.S.: Skill discovery in continuous reinforcement learning domains using skill chaining. In: Advances in Neural Information Processing Systems 22. (2009) 1015–1023
- [19] Niekum, S.: Semantically Grounded Learning from Unstructured Demonstrations. PhD thesis, University of Massachusetts, Amherst (2013)
- [20] Niekum, S.: An integrated system for learning multi-step robotic tasks from unstructured demonstrations. In: AAAI Spring Symposium: Designing Intelligent Robots. (2013)
- [21] Deisenroth, M.P., Fox, D., Rasmussen, C.E.: Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(2) (2015) 408–423
- [22] Deisenroth, M., Rasmussen, C.E.: Pilco: A model-based and data-efficient approach to policy search. In: Proceedings of the 28th International Conference on machine learning (ICML). (2011) 465–472
- [23] Tassa, Y., Mansard, N., Todorov, E.: Control-limited differential dynamic programming. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). (2014) 1168–1175
- [24] Kwakernaak, H., Sivan, R.: Linear optimal control systems. Volume 1. Wiley-interscience, New York (1972)
- [25] Jacobson, D., Mayne, D.: Differential dynamic programming. American Elsevier Pub. Co., New York (1970)
- [26] Mordatch, I., Todorov, E.: Combining the benefits of function approximation and trajectory optimization. In: Robotics: Science and Systems (RSS). (2014)
- [27] McLachlan, G., Peel, D.: Finite mixture models. John Wiley & Sons (2004)
- [28] : Matlab statistics and machine learning toolbox. <http://www.mathworks.com/help/stats/index.html> Accessed: 2015-12-31.
- [29] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)